LAHEDES: the LAGLIDADG homing endonuclease database and engineering server

Gregory K. Taylor[1,2], Lucas H. Petrucci[3], Abigail R. Lambert[4], Sarah K. Baxter[4,5], Jordan Jarjour[6] and Barry L. Stoddard[1,*]

[1]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle WA 98109, [2]Graduate Program in Molecular and Cellular Biology, University of Washington, Seattle WA 98195-7275, [3]Department of Computer Science and Engineering, University of Washington, Seattle WA 98195-2350, [4]Seattle Children's Research Institute, Seattle WA 98101, [5]Medical Scientist Training Program and the Department of Immunology, University of Washington, Seattle WA 98195-7470 and [6]Pregenen Inc. 454 N., Seattle WA 98103, USA

ABSTRACT

LAGLIDADG homing endonucleases (LHEs) are DNA cleaving enzymes, also termed 'meganucleases' that are employed as gene-targeting reagents. This use of LHEs requires that their DNA specificity be altered to match sequences in genomic targets. The choice of the most appropriate LHE to target a particular gene is facilitated by the growing number of such enzymes with well-characterized activities and structures. 'LAHEDES' (The LAGLIDADG Homing Endonuclease Database and Engineering Server) provides both an online archive of LHEs with validated DNA cleavage specificities and DNA-binding interactions, as well as a tool for the identification of DNA sequences that might be targeted by various LHEs. Searches can be performed using four separate scoring algorithms and user-defined choices of LHE scaffolds. The webserver subsequently provides information regarding clusters of amino acids that should be interrogated during engineering and selection experiments. The webserver is fully open access and can be found at http://homingendonuclease.net.

Genome engineering and targeted gene modification is an emerging discipline in which genomes within cell lines, tissues or organisms are manipulated and altered at specified individual loci (1). Such approaches are now being used for a wide-variety of purposes, such as the correction of individual genes in patients suffering from genetic diseases (2); the targeted disruption of genes in patients afflicted by latent viral infections (3); the modification or insertion of genes in plants (4); the generation of unique transgenic stem cell lines (5); the genetic modification of animal model systems (6) and the incorporation of gene drive systems into disease vectors as part of population control strategies (7). By relying on gene-targeting enzymes to enhance the efficiency of site-specific modification, these approaches reduce the burden of intermediate screening steps and use of selectable markers that are necessary in traditional gene-targeting techniques (8–10). A variety of systems are being developed for this purpose, including site-specific recombinases (11) and transposases (12) that directly deliver gene sequences to user-defined chromosomal locations, as well as site-specific DNA cleaving enzymes. These latter systems, ranging from peptide-nucleic acid (PNA) conjugates to site-specific endonucleases, induce double strand breaks that are repaired via homologous recombination and nonhomologous end-joining, resulting in targeted gene modification and disruption, respectively (13).

Three separate protein scaffolds which generate site-specific double-stranded DNA breaks can be used for targeted gene modification: zinc-finger nucleases (ZFNs) (14), TAL effector nucleases (TALENs) (15) and LAGLIDADG homing endonucleases (LHEs; also termed 'Meganucleases') (16). Thus, the field of site-specific genome engineering using site-specific nucleases enjoys a wealth of structural scaffolds for the development of gene-targeting proteins. However, the structures and properties of ZFNS and TALENs differ significantly from LHEs. The first two are artificial chimeras that contain tandem arrays of modular DNA-binding domains tethered to a nonspecific nuclease domain (usually derived from the R.FokI restriction endonuclease). In contrast, LHEs are naturally occurring microbial gene-targeting proteins that are often associated with mobile introns or inteins. The DNA recognition specificities of ZFNs and TALENs are highly engineerable due to their modular

architectures, but both are large proteins that must be introduced and/or expressed as two separately encoded protein chains. In contrast, LHEs are considerably more challenging to redesign, but they possess several properties that otherwise make them ideal candidates for gene targeting, including compact, stable monomeric protein folds that display exceptionally high DNA cleavage specificities (17).

The ability to routinely redirect the DNA target specificity of an LHE is therefore a highly desirable goal. A wide variety of methods have been described for this purpose, including structure-based engineering and several different selection methods [reviewed in (16)]. Until recently, only one LHE (the I-CreI endonuclease) had been successfully engineered and used for the targeted modification of actual physiological coding sequences (18,19), while several additional LHEs (including I-MsoI, I-SceI and I-AniI) had been redesigned in a more limited manner [also reviewed in (16)].

A recent bioinformatics analysis (20) demonstrated that (i) many hundreds of recognizable LHE genes can be found within microbial genomic sequence databases; (ii) a significant fraction of those genes encode active endonucleases; (iii) their target sites can often be determined by a combination of comparative genomic analyses and biochemical experimentation; and (iv) even closely related LHE homologs can display considerably diverged DNA cleavage specificities. These observations were exploited to rapidly generate a novel gene-targeting protein that was highly active against the endogenous human monoamine oxidase B gene (20). Based on those observations, it seems clear that the natural diversity of LHEs can be further exploited in the field of gene targeting, by providing a wealth of initial scaffolds that collectively would reduce the challenge of engineering reagents for individual gene targets. This concept was also suggested in an earlier analysis of putative homing endonuclease genes, which was subsequently organized into the HomeBase webserver (http://homebase-search.tau.ac.il) (21).

The identification and characterization of active LHEs has motivated the development of a dedicated LHE database and webserver, described below, that could serve both as an archive of those LHEs that have been sufficiently characterized to be used as starting points for creating gene-targeting proteins, and as a portal for identifying combinations of endonuclease scaffolds and DNA targets that could be used for targeted gene modification. The resource is not an all-inclusive database of LHEs [of which there are hundreds that are known, most of which are already collected and cataloged in InBase (22), ReBase (23), GISSD (24) and HomeBase (21)] but rather a resource to facilitate the creation of gene-targeting proteins from those LHES that meet at least the three following criteria:

(i) The exact boundaries and center of their DNA target sites are known;
(ii) The orientation and positioning of their N- and C-terminal enzyme domains, relative to the two halves of the DNA target site are known; and

(iii) The identity of amino acid residues that are located within close proximity to clusters of nucleotide bases in the DNA target site are known.

This online resource consists of five tools that are accessible from a top-level group of menu items: an *Endonuclease Browser* that provides the identity and basic properties of well-characterized LHEs and their target sites, *Entry* tools for adding new endonucleases that meet the criteria listed above, their validated target sites and several additional properties, *Genomic Search* tools for the identification of potential target sites that might be modified through the use of wild-type or engineered LHEs, and *Additional Information* menus that provide additional background information both on homing endonuclease biology and on the use of the webserver. The entire resource is designed to be extremely simple and intuitive to use, and instructions for all key steps are provided as online popups that are accessed directly at the individual points of operation for all key functions.

LHE DATABASE

Enzyme and target site entry

The first core functionality of the LAGLIDADG Homing Endonuclease Database and Engineering Server (LAHEDES) resource is an updateable collection of wild-type and engineered LHEs that have been shown to cleave precisely defined DNA sequences corresponding to their target sites in original host gene, and for which the exact positioning and orientation of the target site relative to the LHE's N- and C-terminal domains are known. In the database, the target sites are entered and listed as 22-bp sequences that exactly span the center of four consecutive bases (termed the 'central four target region') that are converted into mutually cohesive four-base 3′-overhangs by all known LHEs. The orientation of the target site relative to the bound endonuclease (i.e. which half-sites are engaged by the N- and C-terminal domains of the LHE, respectively) is defined based on biochemical and/or structural analyses. In this convention, the left half-site (basepairs -11 to -1) is contacted by residues from the N-terminal LHE domain, while the right half-site (basepairs $+1$ to $+11$) is bound by the C-terminal domain.

The 'Entry' tool for addition of new LHEs to the database provides individual fields for the enzyme name, its amino acid sequence (entered as a FASTA character string), its DNA target site (entered from basepair -11 to $+11$ as described above) and a notes field for the entry of additional information for the endonuclease. This information usually includes (i) the host organism and corresponding host gene; (ii) the relationship of the LHE to any surrounding self-splicing elements (introns or inteins) or its existence as a 'stand-alone' endonuclease; (iii) links to genomic or structural information for the endonuclease; and (iv) the nature of experiments that have validated the LHE's activity.

In addition to wild-type LHEs, entry fields are available to archive 'chimeric' LHEs (which correspond to various types of structural fusions between N- and C-terminal LHE domains) (25–27) 'monomerized' LHEs (which correspond to homodimeric endonucleases that have been converted into single-chain reagents through the fusion and tethering of two protein subunits with an artificial peptide linker) (28,29) and fully redesigned LHEs (for those that have been entirely retargeted to genomic target sites and that have been biochemically characterized at a level that matches their wild-type parental enzyme) (20). In each case, the same general entry tools and note fields are provided, with individualized instructions for the information required for adequate archiving. Finally, it is also possible to enter the names, protein sequences and biological origins of 'pseudo-endonucleases', which correspond to LHE genes that do not appear to encode stable or active enzymes (a category motivated by the observation that a significant fraction of homing endonuclease genes have accumulated disabling mutations subsequent to the successful invasion of their host gene) (30).

The rules for naming new LHEs, following the conventions described in Roberts *et al.* (31) are explicitly described and followed by the LAHEDES database, including instructions to examine the REBASE database (23) to ensure that unique acronyms for novel host species and genus are chosen. In general, the entry tool recommends adding an additional character to the end of the acronym to further denote the source of the LHE in cases where a newly discovered enzyme is derived from a mitochondrial ('M') or chloroplast ('C') genome. Certain enzymes that have already enjoyed a long history in the published literature and corresponding databases (such as I-SceI, I-CreI, etc.) maintain their historical names.

As mentioned above, the browser is limited to those LHEs that have been sufficiently characterized that they can be utilized for the creation of gene-targeting reagents (a list that currently spans 24 enzymes). However, the resource is open-ended and available for deposition of any LHEs that meet the necessary criterion for protein engineering and selection experiments.

Specificity profile (position weight matrix) entry

In addition to the exact sequence of their naturally occurring DNA targets (which usually corresponds to the site in the host gene which is cleaved by the homing endonuclease, resulting in invasion of that site by the mobile HE gene and surrounding DNA sequences), the specificity of a given LHE in many cases is also represented in the database by an additional Position Weight Matrix (PWM) (32). This PWM takes into account any positions in the DNA target where base substitutions are tolerated in cleavage experiments (in other words, positions in the DNA target that are recognized with reduced fidelity). This feature of LHEs (the ability to tolerate individual polymorphisms in their natural target sites) is an evolutionary consequence of the need to accommodate the natural drift in the sequence of their host gene targets, and to effectively execute ectopic transfers into new host genes that contain related target sites when the

opportunity arises (30,33). In particular, LHE reading frames that are found within protein-coding genes often tend to display reduced fidelity at positions corresponding to 'wobble' positions in the host gene reading frame (33) [a feature of DNA recognition that is also explicitly accounted for in the HomeBase webserver (21)]. The use of an experimentally determined PWM as a search-scoring matrix exploits this property, but reduces the risk of false positives that might arise from enzymes that depart from this general rule for recognition specificity.

The need to account for relaxed fidelity in LHE–DNA recognition is important in enzyme redesign experiments, where mismatches between the wild-type target site and the desired DNA target which correspond to well-tolerated polymorphisms can be accommodated with minimal effort during the engineering or selection process. Accounting for the specificity profile of an LHE is also important in predicting potential genomic 'off-target' sites, as well as regions of the protein–DNA interface where specificity may be enhanced through additional engineering.

PWMs are entered into the database with a separate entry tool, using information derived either from direct biochemical measurements of the enzyme's ability to cleave alternative target site variants (34) or from experiments that identify collections of cleavable DNA target variants from pools of partially randomized sites (33). In either case, the relative ability of the enzyme to accommodate each possible base at each possible position in the target site is usually entered using a scale of 1.0 (wild-type cleavage activity or recovery for a given base) to 0.0 (complete lack of cleavage activity or recovery for a given base). The resulting PWM values are displayed by the database both as a sequence logo plot, and as raw text that can be incorporated into separate computational analyses (Figure 1).

DNA contact module and redesign mutation entry

In addition to target sites and PWMs, the database accommodates lists of residues for each homing endonuclease that might be potentially useful in any redesign or selection process, and links the information within those lists to the output of target site searches (described below). After entry of a new wild-type LHE, the entry tool for 'contact modules' is then populated by that enzyme in a drop-down menu that allows an investigator to list those amino acids that the investigator believes to be within contact distance to individual clusters of target sites basepairs. This entry tool is arranged for each enzyme into overlapping three base 'modules' across the target site, that can be incorporated into experiments to alter the enzyme's recognition specificity within that same DNA module. For example, the I-OnuI endonuclease displays direct and water-mediated contacts between the −11 to −9 bp positions and at least seven separate amino acid side-chains: Asn 32, Lys 34, Ser 35, Ser 36, Val 37, Gly 38 and Ser 40 (20); all are listed as residues that comprise the 'DNA contact module' for those three corresponding basepairs.
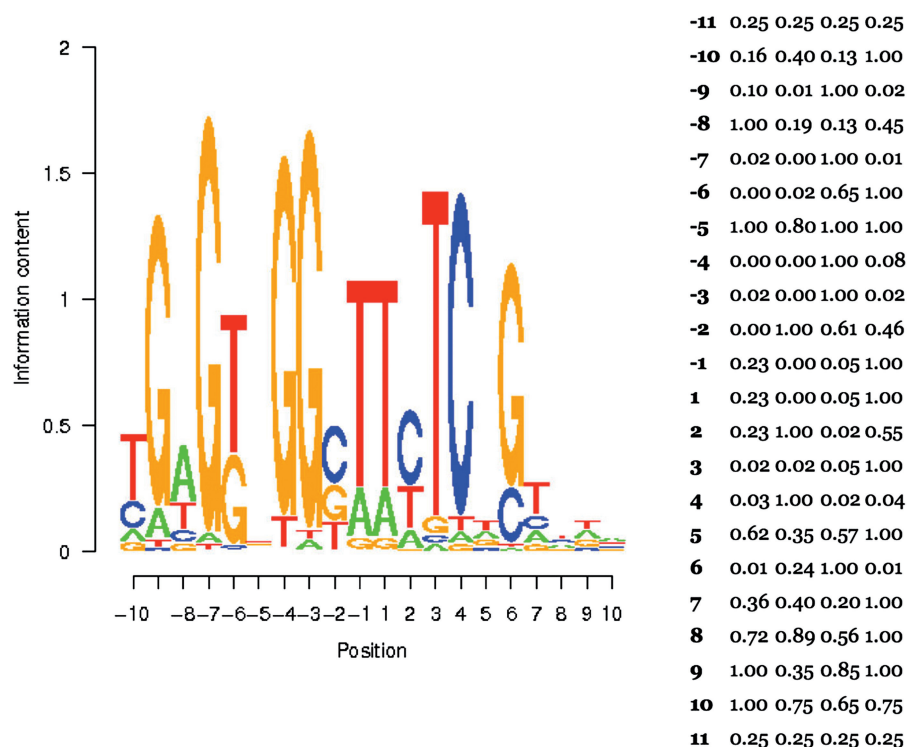
-11	0.25	0.25	0.25	0.25
-10	0.16	0.40	0.13	1.00
-9	0.10	0.01	1.00	0.02
-8	1.00	0.19	0.13	0.45
-7	0.02	0.00	1.00	0.01
-6	0.00	0.02	0.65	1.00
-5	1.00	0.80	1.00	1.00
-4	0.00	0.00	1.00	0.08
-3	0.02	0.00	1.00	0.02
-2	0.00	1.00	0.61	0.46
-1	0.23	0.00	0.05	1.00
1	0.23	0.00	0.05	1.00
2	0.23	1.00	0.02	0.55
3	0.02	0.02	0.05	1.00
4	0.03	1.00	0.02	0.04
5	0.62	0.35	0.57	1.00
6	0.01	0.24	1.00	0.01
7	0.36	0.40	0.20	1.00
8	0.72	0.89	0.56	1.00
9	1.00	0.35	0.85	1.00
10	1.00	0.75	0.65	0.75
11	0.25	0.25	0.25	0.25

Figure 1. The specificity profile (PWM) for the I-AniI LHE. Specificity profiles, reported as PWMs that have been experimentally determined for a number of LHEs can be entered individually and are available on the site as part of the browser functionality. These matrices are displayed in graphical format as a sequence logo plot (left) and in tabular format. The ability of the enzyme to cleave DNA targets that contain a given nucleotide at a specific position within the target site is indicated by letter height, and positions with greater overall information content are more specifically recognized. The profile displayed in this figure was determined in a series of biochemical experiments where the cleavability of target site variants was individually measured (34).

The choice of amino acids that should populate individual DNA contact modules can be derived either from crystallographic structures of a corresponding LHE–DNA complex, or from homology models constructed from such available structures. The contact residues entered into the database for the I-CreI homodimeric LHE and the I-OnuI monomeric LHE have been extensively validated for engineering purposes, and those can be used as an initial reference for the entry of corresponding residue lists for homologous LHEs. The premise behind entering amino acid contact lists in a modular format is based on past experimental observations that sequence recognition within the LHE–DNA interface often displays considerable context dependence and cross-talk between adjacent positions (28,34,35), such that even when only 1-bp substitution is required in the novel target site, best practice often dictates that residues involved in contacts to the '$n-1$' and '$n+1$' basepairs should also be subjected to redesign or selection.

For many enzymes in the database, specific mutations in the DNA contacting residues described above have been shown to be associated with defined alterations in target sequence specificity (20,34,36). For example, incorporation of three amino acid substitutions in the monomerized version of I-MsoI (I30E, S43R and I85Y) results in a shift in specificity from a wild-type A:T basepair at position −8 (termed '−8A' in the database) to preferential recognition of a G:C basepair at the same

position ('−8G') (35). In addition to the entry tool for wild-type DNA contacting residues, an additional 'Specificity Changing Mutation Entry' tool is provided that allows the archiving of such substitutions in the protein scaffold and the corresponding change in DNA base preference in the altered target site. Similar to the list of wild-type DNA contacting residues, the identities of such mutations are provided to the user both in the endonuclease browser tool, and as a subsequent link from the output of the genomic sequence target site searches to aid in the design of engineering or selection experiments.

LHE SEARCH AND ENGINEERING SERVER

The second core functionality of the LAHEDES resource is the ability to search individual genes or collections of genes for potential matches against LHE target sites based on four separate search criteria:

(i) Searches for DNA sequences that exhibit identity across the central 4 bp of an LHE target site, because those bases are usually not in direct contact with amino acid side chains. Therefore, specificity changes at these basepairs are often not easily achieved through protein engineering or selection (34). Additional experiments on individual LHE enzymes that further profile the ability of the

enzyme to accommodate individual basepair substitutions in the 'central four' region of the target site expands the number of hits within a genomic query. For those enzymes where such information is available, the 'Central Four' search algorithm takes such polymorphisms into account when returning potential targets and allows such sites to appear on the list of target site hits.

(ii) Searches for highest possible sequence identity (fewest DNA basepair mismatches) across a 22-bp DNA sequence in the genomic query relative to an LHE's entire physiological target site. Some redesign experiments indicate, not surprisingly, that the most effective way to maintain wild-type levels of affinity, activity and overall specificity is to maintain the highest degree of sequence similarity to the wild-type target (20).

(iii) For those LHEs that have had their complete specificity profile determined, the user is provided with the opportunity to search and score potential targets using a 'Specificity Profile PWM' described above. In this search algorithm, the penalty for a mismatch between a protein's natural target site and a potential genomic target is differentially weighted depending on the fidelity of recognition displayed at each DNA basepair position. The difference between the two search strategies described in (ii) versus (iii) is simple: the use of a simple identity matrix will return those potential targets with the fewest mismatches relative to the enzyme's wild-type recognition site, while the use of a scoring matrix that accounts for recognition degeneracy would indicate sites that might be more distantly related to the protein's wild-type recognition site, while nevertheless indicating more tractable gene-targeting sites.

(iv) The I-OnuI LHE has been systematically tested for its ability to be retargeted towards each possible altered base triplet across the entire length of the

(a) I-OnuI

Gene	Position	-11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11	Score
>gi|156630984|ref|NM_000898.4|	r 2270/2611	g g T C C A C a T A T Tt A A C C T T T T g	3.11348e-08
>gi|156630984|ref|NM_000898.4|	f 2079/2611	T T T g t g C T T g T T C t t C C T c T T A	3.11348e-10
>gi|156630984|ref|NM_000898.4|	r 1652/2611	T a a t C A C a T A T Tt A g C C T c a T A	3.11348e-11
>gi|156630984|ref|NM_000898.4|	r 36/2611	T T T t t A a T T A c c C A A g t g T g T A	3.11348e-12
>gi|156630984|ref|NM_000898.4|	f 2501/2611	a T T t t A t c T c T T C t t C C T T T T g	3.11348e-12

(b) I-OnuI

Gene	Position	-11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 1 2 3 4 5 6 7 8 9 10 11	Score
wild type sequence	N/A	T T T C C A C T T A T T C A A C C T T T T A	2.52e+08
>gi|156630984|ref|NM_000898.4|	f 957/2611	a T T a g t g c T A T T C c t C C T a c T c	1.3608e+08
>gi|156630984|ref|NM_000898.4|	r 1697/2611	c a a g g A C a T t T T C t c t t g T c T g	9.3312e+07
>gi|156630984|ref|NM_000898.4|	r 2340/2611	a a c a a c C a c A T T C A g t C c a g a g	6.86e+07
>gi|156630984|ref|NM_000898.4|	r 1534/2611	T a a t a A a c T A T a C A c t t g a T g A	6.77376e+07
>gi|156630984|ref|NM_000898.4|	r 2471/2611	c T g C t c C g T t T T C t g g g c c T c g	5.6448e+07

(c)

(d) Amino Acid Residues for I-OnuI | Amino Acid Residues for I-OnuI

Module -11 → -9: | Module -8 → -6:

Asn 32, Lys 34, Ser 35, Ser 36, Val 37, Gly 38, Ser 40 | Leu 26, Arg 28, Arg 30, Glu 42, Gly 44, Val 68, Ala 70, Lys 80, Thr 82

Figure 2. Genomic search results. The human monoamine oxidase B (MAOB) gene was searched for potential target sites matches using the I-OnuI LHE and either the 'Identity' algorithm (**a**) or the 'Modular Engineerability' algorithm (**b**). For both searches, a list of best matches is shown, together with their position in the query sequence, orientation (forward or reverse) and scores. The top scoring hit in the 'Identity' Search (Panel A, top MAOB sequence) was targeted for cleavage at the endogenous chromosomal MAOB locus by an engineered version of I-OnuI (20) (listed as 'I-OnuI(E2)' in the endonuclease browser, see Figure 1). Throughout the target site hits, mismatches are shown as lower base bases. In the output from a Module Search, each score is linked to a more detailed view (**c**) illustrating how well each DNA module scores using an 'engineerability' matrix for I-OnuI. Modules are represented by individual bars; those that score favorably in a given target are colored blue, while those that do not score as favorably are colored red. Nucleotides that directly match the wild type target are indicated by vertical bars. The engineerability scoring scale (0 to 10; 10 is best) is shown to the right. Clicking on individual bases or modular bars in the search output provides links to lists of amino acids (**d**) and (if they are available) mutations in the protein that are associated with recognition at those positions in the target.

DNA target site, using a high throughput method that combines yeast surface display and flow cytometry (37) (J. Jarjour, unpublished data). For that enzyme, genomic searches are therefore possible using corresponding modular scoring matrices, returning lists of potential target sites to the investigator that might be accessible for modification through a subsequent experimental strategy that makes use of highthroughput selection from pools of enzyme variants that are randomized at corresponding DNA-contacting residues.

The search algorithm also allows the user to search for genomic sequences that might be targeted through the application of validated 'chimeric' homing endonucleases (where N- and C-terminal domains or DNA contacting surfaces of unrelated LHEs have been fused or individually modified to create scaffolds that can recognize corresponding chimeric DNA target sites) (25–27).

Following genomic target site searches using any of the four methods outlined above (along with user-chosen endonucleases for the search that can span all or a smaller subset of those LHEs that are available in the database), a top-scoring list of putative target sequences is returned in a graphical user interface, alongside the enzyme's wild-type target for comparison (Figure 2a). The output, which can be saved to a text file, indicates the position and identity of all basepairs in the potential target that differs from the enzyme's wild-type recognition site, and provides links from each base in the candidate targets to information that has been entered about amino acids and mutations in the LHE that might influence specificity and redesign at that position.

While the search output using the first three scoring algorithms ('Central 4 Match', 'Target Identity' and 'PWM Search') is arranged to clearly display the conservation or mismatch at each individual DNA basepair, the output for the 'Modular' search displays individual genomic sequences in the form of information regarding the 'engineerability' of each protein–DNA module that harbors one or more basepair changes (Figure 2b). As described above, the individual regions and basepairs in potential target sites are linked to previously entered information regarding the identity of amino acids and the corresponding mutations that are involved in the recognition of those positions.

CONCLUSIONS

The field of genome and gene engineering is currently experiencing a rapid increase in the amount of information and data relating to the creation of gene-targeting proteins derived from the LHE family. These include: (i) the continuing identification of new LHEs from worldwide microbial and metagenomic sequencing projects and (ii) the generation of a large numbers of engineered LHE variants with altered DNA recognition properties. One manner in which the full potential and impact of this information can be fully exploited is through the development and maintenance of a computational database and search engine for LHE proteins.

Although LHEs have historically been considerably more difficult to engineer for altered DNA recognition and cleavage properties relative to their artificial, modular ZFN and TALEN counterparts, there still exists considerable industrial and academic activity and motivation to reduce or eliminate the technical gulf that separates them from routine application in genome engineering applications. It is likely that the combination of highthroughput modular screens for altered specificity and the increasing numbers of wild-type and chimeric LHE protein scaffolds, along with the ability to apply deep sequencing methods to resulting pools of highly active enzyme variants, may soon drive a significant reduction in the time and cost of assembling artificial LHEs for the life science community.

REFERENCES

1. McMahon,M.A., Rahdar,M. and Porteus,M. (2012) Gene editing: not just for translation anymore. *Nat. Methods*, **9**, 28–31.
2. Ellis,B.L., Hirsch,M.L., Porter,S.N., Samulski,R.J. and Porteus,M.H. (2012) Zinc-finger nuclease-mediated gene correction using single AAV vector transduction and enhancement by Food and Drug Administration-approved drugs. *Gene Ther.*, January 19 (doi:10.1038/gt.2011.211; epub ahead of print).
3. Cannon,P. and June,C. (2011) Chemokine receptor 5 knockout strategies. *Curr. Opin. HIV AIDS*, **6**, 74–79.
4. Kim,S. and Kim,J.S. (2011) Targeted genome engineering via zinc finger nucleases. *Plant Biotechnol. Rep.*, **5**, 9–17.
5. Osiak,A., Radecke,F., Guhl,E., Radecke,S., Dannemann,N., Lutge,F., Glage,S., Rudolph,C., Cantz,T., Schwarz,K. *et al.* (2011) Selection-independent generation of gene knockout mouse embryonic stem cells using zinc-finger nucleases. *PloS One*, **6**, e28911.
6. Ekker,S.C. (2008) Zinc finger-based knockout punches for zebrafish genes. *Zebrafish*, **5**, 121–123.
7. Windbichler,N., Menichelli,M., Papathanos,P.A., Thyme,S.B., Li,H., Ulge,U.Y., Hovde,B.T., Baker,D., Monnat,R.J. Jr, Burt,A. *et al.* (2011) A synthetic homing endonuclease-based gene drive system in the human malaria mosquito. *Nature*, **473**, 212–215.
8. Glaser,S., Anastassiadis,K. and Stewart,A.F. (2005) Current issues in mouse genome engineering. *Nat. Genet.*, **37**, 1187–1193.
9. Tzfira,T. and White,C. (2005) Towards targeted mutagenesis and gene replacement in plants. *Trends Biotechnol.*, **23**, 567–569.
10. Posfai,G., Kolisnychenko,V., Bereczki,Z. and Blattner,F.R. (1999) Markerless gene replacement in Escherichia coli stimulated by a double-strand break in the chromosome. *Nucleic Acids Res.*, **27**, 4409–4415.
11. Gersbach,C.A., Gaj,T., Gordley,R.M., Mercer,A.C. and Barbas,C.F. 3rd (2011) Targeted plasmid integration into the human genome by an engineered zinc-finger recombinase. *Nucleic Acids Res.*, **39**, 7868–7878.
12. Kettlun,C., Galvan,D.L., George,A.L. Jr, Kaja,A. and Wilson,M.H. (2011) Manipulating piggyBac transposon

chromosomal integration site selection in human cells. *Mol. Ther. J. Am. Soc. Gene Ther.*, **19**, 1636–1644.

13. Aiba,Y., Sumaoka,J. and Komiyama,M. (2011) Artificial DNA cutters for DNA manipulation and genome engineering. *Chem. Soc. Rev.*, **40**, 5657–5668.

14. Carroll,D. (2011) Genome engineering with zinc-finger nucleases. *Genetics*, **188**, 773–782.

15. Bogdanove,A.J. and Voytas,D.F. (2011) TAL effectors: customizable proteins for DNA targeting. *Science*, **333**, 1843–1846.

16. Stoddard,B.L. (2011) Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. *Structure*, **19**, 7–15.

17. Paques,F. and Duchateau,P. (2007) Meganucleases and DNA double-strand break-induced recombination: perspectives for gene therapy. *Curr. Gene Ther.*, **7**, 49–66.

18. Gao,H., Smith,J., Yang,M., Jones,S., Djukanovic,V., Nicholson,M.G., West,A., Bidney,D., Falco,S.C., Jantz,D. *et al.* (2010) Heritable targeted mutagenesis in maize using a designed endonuclease. *Plant J Cell Mol. Biol.*, **61**, 176–187.

19. Munoz,I.G., Prieto,J., Subramanian,S., Coloma,J., Redondo,P., Villate,M., Merino,N., Marenchino,M., D'Abramo,M., Gervasio,F.L. *et al.* (2011) Molecular basis of engineered meganuclease targeting of the endogenous human RAG1 locus. *Nucleic Acids Res.*, **39**, 729–743.

20. Takeuchi,R., Lambert,A.R., Mak,A.N., Jacoby,K., Dickson,R.J., Gloor,G.B., Scharenberg,A.M., Edgell,D.R. and Stoddard,B.L. (2011) Tapping natural reservoirs of homing endonucleases for targeted gene modification. *Proc. Natl Acad. Sci. USA*, **108**, 13077–13082.

21. Barzel,A., Privman,E., Peeri,M., Naor,A., Shachar,E., Burstein,D., Lazary,R., Gophna,U., Pupko,T. and Kupiec,M. (2011) Native homing endonucleases can target conserved genes in humans and in animal models. *Nucleic Acids Res.*, **39**, 6646–6659.

22. Perler,F.B. (2002) InBase: the Intein Database. *Nucleic Acids Res.*, **30**, 383–384.

23. Roberts,R.J., Vincze,T., Posfai,J. and Macelis,D. (2010) REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Res.*, **38**, D234–D236.

24. Zhou,Y., Lu,C., Wu,Q.J., Wang,Y., Sun,Z.T., Deng,J.C. and Zhang,Y. (2008) GISSD: Group I Intron Sequence and Structure Database. *Nucleic Acids Res.*, **36**, D31–D37.

25. Chevalier,B.S., Kortemme,T., Chadsey,M.S., Baker,D., Monnat,R.J. and Stoddard,B.L. (2002) Design, activity, and structure of a highly specific artificial endonuclease. *Mol. Cell*, **10**, 895–905.

26. Grizot,S., Epinat,J.C., Thomas,S., Duclert,A., Rolland,S., Paques,F. and Duchateau,P. (2010) Generation of redesigned homing endonucleases comprising DNA-binding domains derived from two different scaffolds. *Nucleic Acids Res.*, **38**, 2006–2018.

27. Silva,G.H., Belfort,M., Wende,W. and Pingoud,A. (2006) From monomeric to homodimeric endonucleases and back: engineering novel specificity of LAGLIDADG enzymes. *J. Mol. Biol.*, **361**, 744–754.

28. Grizot,S., Smith,J., Daboussi,F., Prieto,J., Redondo,P., Merino,N., Villate,M., Thomas,S., Lemaire,L., Montoya,G. *et al.* (2009) Efficient targeting of a SCID gene by an engineered single-chain homing endonuclease. *Nucleic Acids Res.*, **37**, 5405–5419.

29. Li,H., Pellenz,S., Ulge,U., Stoddard,B.L. and Monnat,R.J. Jr (2009) Generation of single-chain LAGLIDADG homing endonucleases from native homodimeric precursor proteins. *Nucleic Acids Res.*, **37**, 1650–1662.

30. Posey,K.L., Koufopanou,V., Burt,A. and Gimble,F.S. (2004) Evolution of divergent DNA recognition specificities in VDE homing endonucleases from two yeast species. *Nucleic Acids Res.*, **32**, 3947–3956.

31. Roberts,R.J., Belfort,M., Bestor,T., Bhagwat,A.S., Bickle,T.A., Bitinaite,J., Blumenthal,R.M., Degtyarev,S., Dryden,D.T., Dybvig,K. *et al.* (2003) A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.*, **31**, 1805–1812.

32. Henikoff,S. (1996) Scores for sequence searches and alignments. *Curr. Opin. Struct. Biol.*, **6**, 353–360.

33. Scalley-Kim,M., McConnell-Smith,A. and Stoddard,B.L. (2007) Coevolution of a homing endonuclease and its host target sequence. *J. Mol. Biol.*, **372**, 1305–1319.

34. Thyme,S.B., Jarjour,J., Takeuchi,R., Havranek,J.J., Ashworth,J., Scharenberg,A.M., Stoddard,B.L. and Baker,D. (2009) Exploitation of binding energy for catalysis and design. *Nature*, **461**, 1300–1304.

35. Ashworth,J., Taylor,G.K., Havranek,J.J., Quadri,S.A., Stoddard,B.L. and Baker,D. (2010) Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Res.*, **38**, 5601–5608.

36. Ashworth,J., Havranek,J.J., Duarte,C.M., Sussman,D., Monnat,R.J. Jr, Stoddard,B.L. and Baker,D. (2006) Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, **441**, 656–659.

37. Jarjour,J., West-Foyle,H., Certo,M.T., Hubert,C.G., Doyle,L., Getz,M.M., Stoddard,B.L. and Scharenberg,A.M. (2009) High-resolution profiling of homing endonuclease binding and catalytic specificity using yeast surface display. *Nucleic Acids Res.*, **37**, 6871–6880.